

Poisson convolution on a tree of categories for modeling topical content with word frequency and exclusivity *

Jonathan M. Bischof, Edoardo M. Airoldi
Department of Statistics
Harvard University, Cambridge, MA 02138, USA

* Address correspondence to EM Airoldi, airoldi@fas.harvard.edu.

Abstract

An ongoing challenge in the analysis of document collections is how to summarize content in terms of a set of inferred *themes* that can be interpreted substantively in terms of topics. However, the current practice of parameterizing the themes in terms of most frequent words limits interpretability by ignoring the differential use of words across topics. We argue that words that are both common and exclusive to a theme are more effective at characterizing topical content. We consider a setting where professional editors have annotated documents to a collection of topic categories, organized into a tree, in which leaf-nodes correspond to the most specific topics. Each document is annotated to multiple categories, possibly at different levels of the tree. We introduce Hierarchical Poisson Convolution (HPC) as a model to analyze annotated documents in this setting. The model leverages the structure among categories defined by professional editors to infer a clear semantic description for each topic in terms of words that are both frequent and exclusive. We develop a parallelized Hamiltonian Monte Carlo sampler that allows the inference to scale to millions of documents.

Keywords: High-dimensional Data; Categorical Data; Hamiltonian Monte Carlo; Parallel Inference; Text Analysis

Contents

1	Introduction	4
2	Hierarchical Poisson Convolution	6
2.1	Modeling word usage rates on the hierarchy	6
2.2	Modeling the topic membership of documents	8
2.3	Estimands	8
3	Scalable inference via parallelized HMC sampler	9
3.1	Block Gibbs Sampler	10
3.1.1	Updating tree parameters	10
3.1.2	Updating topic affinity parameters	11
3.1.3	Updating corpus-level parameters	11
3.2	Estimation	12
3.3	Inference for unlabeled documents	13
4	Results	13
4.1	The Reuters Corpus dataset	13
4.2	How the differential usage parameters regulate topic exclusivity	16
4.3	How frequency modulates regularization of exclusivity	17
4.4	Frequency and Exclusivity as a two dimensional summary of semantic content	19
4.5	Classification performance	19
5	Discussion	23
5.1	Concluding remarks	24
A	Appendix: Implementing the parallelized HMC sampler	28
A.1	Hamiltonian Monte Carlo conditional updates	28
A.2	SCHMC implementation details for HPC model	29
A.2.1	Conditional posterior of the rate parameters	29
A.2.2	Conditional posterior of the topic affinity parameters	32
A.2.3	Conditional posterior of the τ_{fk}^2 hyperparameters	33

1 Introduction

A recurrent challenge in the multivariate statistics is how to construct interpretable low-dimensional summaries of high-dimensional data. Historically, simple models based on correlation matrices, such as principal component analysis (Jolliffe, 1986) and canonical correlation analysis (Hotelling, 1936), have proven to be effective tools for data reduction. More recently, multilevel models have become a flexible and powerful tool for finding latent structure in high dimensional data (McLachlan and Peel, 2000; Sohn and Xing, 2009; Blei et al., 2003b; Airoldi et al., 2008). However, while interpretable statistical summaries are highly valued in applications, dimensionality reduction models are rarely optimized to aid qualitative discovery; there is no guarantee that the optimal low-dimensional projections will be understandable in terms of quantities of scientific interest that can help practitioners make decisions. Instead, we design a model with scientific estimands of interest in mind to achieve an optimal balance of interpretability and dimensionality reduction.

We consider a setting in which we observe two sets of categorical data for each unit of observation: $\mathbf{w}_{1:V}$, which live in a high-dimensional space, and $\mathbf{l}_{1:K}$, which live in a structured low-dimensional space and provide a direct link to information of scientific interest about the sampling units. The goal of the analysis is two fold. First, we desire to develop a joint model for the observations $\mathbf{Y} \equiv \{\mathbf{W}_{D \times V}, \mathbf{L}_{D \times K}\}$ that can be used to project the data onto a low-dimensional parameter space Θ in which interpretability is maintained by mapping categories in \mathcal{L} to directions in Θ . Second, we would like the mapping from the original space to the low-dimensional projection to be scientifically interesting so that statistical insights about Θ can be understood in terms of the original inputs, $\mathbf{w}_{1:V}$, in a way that guides future research.

In the application to text analysis that motivates this work, $\mathbf{w}_{1:N}$ are the raw word counts observed in each document and $\mathbf{l}_{1:K}$ are a set of labels created by professional editors that are indicative of topical content. Specifically, the words are represented as an unordered vector of counts, with the length of the vector corresponding to the size of a known dictionary. The labels are organized in a tree-structured ontology, from the most generic topic at the root of the tree to the most specific topic at the leaves. Each news article may be annotated with more than one label, at the editors’ discretion. The number of labels is given by the size of the ontology and typically ranges from tens to hundreds of categories. In this context, the inferential challenge is to discover a low dimensional representation of topical content, Θ , that aligns with the coarse labels provided by editors while at the same time providing a mapping between the textual content and directions in Θ in a way that formalizes and enhances our understanding of how low dimensional structure is expressed the space of observed words.

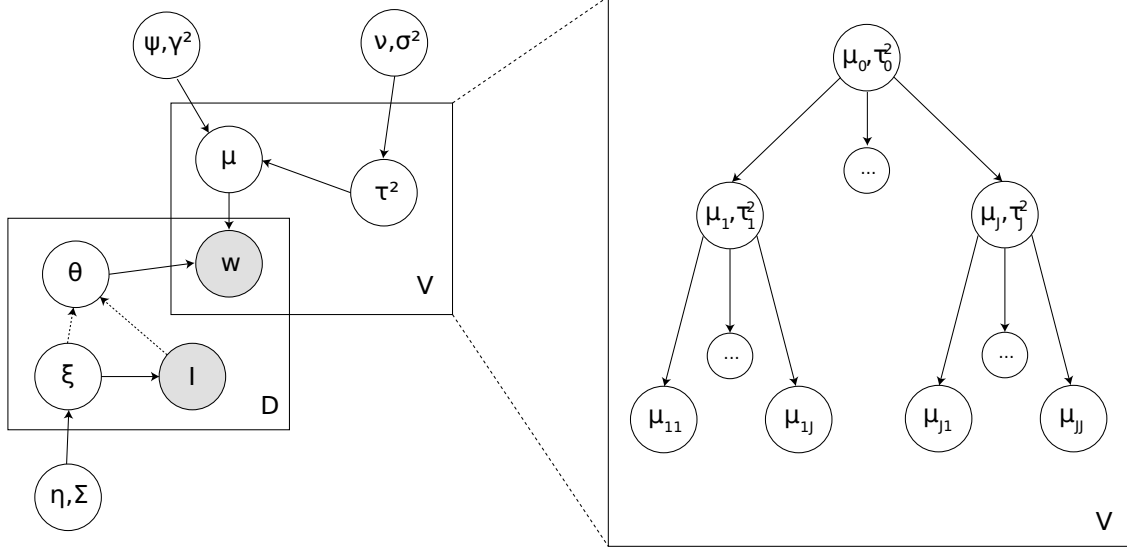
Recent approaches to this problem in the machine learning literature have taken a Bayesian hierarchical approach to this task by viewing a document’s content as arising from a mixture of component distributions, commonly referred to as “topics” as they often capture thematic structure (Blei, 2012). As the component distributions are almost exclusively parameterized as multinomial distributions over words in the vocabulary, the loading of words onto topics is characterized in terms of the relative frequency of within-component usage. While relative frequency has proven to be a useful mapping of topical content onto words, recent work has documented a growing list of interpretability issues with frequency-based summaries: they are often dominated by countless

“stop” words (Wallach et al., 2009), sometimes appear incoherent or redundant (Mimno et al., 2011; Chang et al., 2009; Airoldi et al., 2010), and typically require post hoc modification to meet human expectations (Hu et al., 2011; Grimmer and King, 2011). Instead, we propose a new mapping for topical content that incorporates how words are used differentially across topics. If a word is common in a topic, it is also important to know whether it is common in many topics or relatively exclusive to the topic in question. Both of these summary statistics are informative: nonexclusive words are less likely to carry topic-specific content, while infrequent words occur too rarely to form the semantic core of a topic. We therefore look for the most frequent words in the corpus that are also likely to have been generated from the topic of interest to summarize its content. In this approach we borrow ideas from the statistical literature, in which models of differential word usage have been leveraged for analyzing writing styles in a supervised setting (Mosteller and Wallace, 1984; Airoldi et al., 2005, 2006, 2007; Monroe et al., 2008), and combine them with ideas from the machine learning literature, in which latent variable and mixture models based on frequent word usage have been used to infer structure that often captures topical content (McCallum et al., 1998; Blei et al., 2003b; Canny, 2004).

From a statistical perspective, models based on topic-specific distributions over the vocabulary cannot produce stable estimates of differential usage since they only model the relative frequency of words within topics. They cannot regularize usage across topics and naively infer the greatest differential usage for the rarest features (Eisenstein et al., 2011). To tackle this issue, we introduce the generative framework of Hierarchical Poisson Convolution (HPC) that parameterizes topic-specific word counts as unnormalized count variates whose rates can be regularized across topics as well as within them, making stable inference of both word frequency and exclusivity possible. HPC can be seen as a fully generative extension of Sparse Topic Coding (Zhu and Xing, 2011) that emphasizes regularization and interpretability rather than exact sparsity. Additionally, HPC leverages hierarchical systems of topic categories created by professional editors in collections such as *Reuters*, *New York Times*, *Wikipedia*, and *Encyclopedia Britannica* to make focused comparisons of differential use between neighboring topics on the tree and build a sophisticated joint model for topic memberships and labels in the documents. By conditioning on a known hierarchy, we avoid the complicated task of inferring hierarchical structure (Blei et al., 2003a; Mimno et al., 2007; Adams et al., 2010). We introduce a parallelized Hamiltonian Monte Carlo (HMC) estimation strategy that makes full Bayesian inference efficient and scalable.

The proposed model is designed to infer an interpretable description of human-generated labels, thus we restrict the topic components to have a one-to-one correspondence with the human-generated labels, as in Labeled LDA (Ramage et al., 2009). This *descriptive* link between the labels and topics differs from the *predictive* link used in Supervised LDA (Blei and McAuliffe, 2007; Perotte et al., 2012), where topics are learned as an optimal covariate space to predict an observed document label or response variable. The more restrictive descriptive link can be expected to limit predictive power, but is crucial for learning summaries of individual labels. We then infer a description of these labels in terms of words that are both frequent and exclusive. We anticipate that learning a concise semantic description for any collection of topics implicitly defined by professional editors is the first step toward the semi-automated creation of domain-specific topic ontologies. Domain-specific topic ontologies may be useful for evaluating the semantic content of *inferred* topics, or for predicting the semantic content of new social media, including Twitter

Figure 1: Graphical representation of Hierarchical Poisson Convolution (left) and detail on tree plate (right)



messages and Facebook wall-posts.

2 Hierarchical Poisson Convolution

The Hierarchical Poisson Convolution model is a data generating process for document collections whose topics are organized in a hierarchy, and whose topic labels are observed. We refer to the structure among topics interchangeably as a *hierarchy* or *tree* since we assume that each topic has exactly one parent and that no cyclical parental relations are allowed. Each document $d \in \{1, \dots, D\}$ is a record of counts w_{fd} for every feature in the vocabulary, $f \in \{1, \dots, V\}$. The length of the document is given by L_d , which we normalize by the average document length L to get $l_d \equiv \frac{1}{L}L_d$. Documents have unrestricted membership to any combination of topics $k \in \{1, \dots, K\}$ represented by a vector of labels \mathbf{I}_d where $I_{dk} \equiv I\{\text{doc } d \text{ belongs to topic } k\}$.

2.1 Modeling word usage rates on the hierarchy

The HPC model leverages the known topic hierarchy by assuming that words are used similarly in neighboring topics. Specifically, the log rate for a word across topics follows a Gaussian diffusion down the tree. Consider the topic hierarchy presented in the right panel of Figure 1. At the top level, $\mu_{f,0}$ represents the log rate for feature f overall in the corpus. The log rates $\mu_{f,1}, \dots, \mu_{f,J}$ for first level topics are then drawn from a Gaussian centered around the corpus rate with dispersion controlled by the variance parameter $\tau_{f,0}^2$. From first level topics, we then draw the log rates for the second level topics from another Gaussian centered around their mean $\mu_{f,j}$ and with variance

Table 1: Generative process for Hierarchical Poisson Convolution

Step	Generative process
Tree parameters	For feature $f \in \{1, \dots, V\}$: <ul style="list-style-type: none"> • Draw $\mu_{f,0} \sim \mathcal{N}(\psi, \gamma^2)$ • Draw $\tau_{f,0}^2 \sim \text{Scaled Inv-}\chi^2(\nu, \sigma^2)$ • For $j \in \{1, \dots, J\}$ (first level of hierarchy): <ul style="list-style-type: none"> – Draw $\mu_{f,j} \sim \mathcal{N}(\mu_{f,0}, \tau_{f,0}^2)$ – Draw $\tau_{f,j}^2 \sim \text{Scaled Inv-}\chi^2(\nu, \sigma^2)$ • For $j \in \{1, \dots, J\}$ (terminal level of hierarchy): <ul style="list-style-type: none"> – Draw $\mu_{f,j1}, \dots, \mu_{f,jJ} \sim \mathcal{N}(\mu_{f,j}, \tau_{f,j}^2)$ • Define $\beta_{f,k} \equiv e^{\mu_{f,k}}$ for $k \in \{1, \dots, K\}$
Topic membership parameters	For document $d \in \{1, \dots, D\}$: <ul style="list-style-type: none"> • Draw $\xi_d \sim \mathcal{N}(\eta, \Sigma = \lambda^2 \mathbf{I}_K)$ • For topic $k \in \{1, \dots, K\}$: <ul style="list-style-type: none"> – Define $p_{dk} \equiv 1/(1 + e^{-\xi_{dk}})$ – Draw $I_{dk} \sim \text{Bernoulli}(p_{dk})$ – Define $\theta_{dk}(\mathbf{I}_d, \xi_d) \equiv e^{\xi_{dk}} I_{dk} / \sum_{j=1}^K e^{\xi_{dj}} I_{dj}$
Data generation	For document $d \in \{1, \dots, D\}$: <ul style="list-style-type: none"> • Draw normalized document length $l_d \sim \frac{1}{L} \text{Pois}(v)$ • For every topic k and feature f: <ul style="list-style-type: none"> – Draw count $w_{fdk} \sim \text{Pois}(l_d \theta_d^T \beta_f)$ • Define $w_{fd} \equiv \sum_{k=1}^K w_{fdk}$ (observed data)

$\tau_{f,j}^2$. This process is continued down the tree, with each parent node having a separate variance parameter to control the dispersion of its children.

The variance parameters τ_{fp}^2 directly control the local differential expression in a branch of the tree. Words with high variance parameters can have rates in the child topics that differ greatly from the parent topic p , allowing the child rates to diverge. Words with low variance parameters will have rates close to the parent and so will be expressed similarly among the children. If we learn a population distribution for the τ_{fp}^2 that has low mean and variance, it is equivalent to saying that most features are expressed similarly across topics *a priori* and that we would need a preponderance of evidence to believe otherwise.

2.2 Modeling the topic membership of documents

Documents in the HPC model can contain content from any of the K topics in the hierarchy at varying proportions, with the exact allocation given by the vector θ_d on the $K - 1$ simplex. The model assumes that the count for word f contributed by each topic follows a Poisson distribution whose rate is moderated by the document’s length and membership to the topic; that is, $w_{fdk} \sim \text{Pois}(l_d \theta_{dk} \beta_{fk})$. The only data we observe is the total word count $w_{fd} \equiv \sum_{k=1}^K w_{fdk}$, but the infinite divisibility property of the Poisson distribution gives us that $w_{fd} \sim \text{Pois}(l_d \theta_d^T \beta_f)$. These draws are done for every word in the vocabulary (using the same θ_d) to get the content of the document.¹

In labeled document collections, human coders give us an extra piece of information for each document, I_d , that indicates the set of topics that contributed its content. As a result, we know $\theta_{dk} = 0$ for all topics k where $I_{dk} = 0$, and only have to determine how content is allocated between the set of active topics.

The HPC model assumes that these two sources of information for a document are not generated independently. A document should not have a high probability of being labeled to a topic from which it receives little content and vice versa. Instead, the model posits a latent K -dimensional topic affinity vector $\xi_d \sim \mathcal{N}(\eta, \Sigma)$ that expresses how strongly the document is associated with each topic. The topic memberships and labels of the document are different manifestations of this affinity. Specifically, each ξ_{dk} is the log odds that topic label k is active in the document, with $I_{dk} \sim \text{Bernoulli}(\text{logit}^{-1}(\xi_{dk}))$. Conditional on the labels, the topic memberships are the relative sizes of the document’s affinity for the active topics and zero for inactive topics: $\theta_{dk} \equiv e^{\xi_{dk}} I_{dk} / \sum_{j=1}^K e^{\xi_{dj}} I_{dj}$. Restricting each document’s membership vectors to the labeled topics is a natural and efficient way to generate sparsity in the mixing parameters, stabilizing inference and reducing the computational burden of posterior simulation.

We outline the generative process in full detail in Table 1, which can be summarized in three steps. First, a set of rate and variance parameters are drawn for each feature in the vocabulary. Second, a topic affinity vector is drawn for each document in the corpus, which generate topic labels. Finally, both sets of parameters are then used to generate the words in each document. For simplicity of presentation we assume that each non-terminal node has J children and that the tree has only two levels below the corpus level, but the model can accommodate any tree structure.

2.3 Estimands

In order to measure topical semantic content, we consider the topic-specific frequency and exclusivity of each word in the vocabulary. These quantities form a two-dimensional summary of each word’s relation to a topic of interest, with higher scores in both being positively related to topic specific content. Additionally, we develop a univariate summary of semantic content that can be used to rank words in terms of their semantic content. These estimands are simple functions of the rate parameters of HPC; the distribution of the documents’ topic memberships is a nuisance parameter needed to disambiguate the content of a document between its labeled topics.

¹This is where the model’s name arises: the observed feature count in each document is the convolution of (unobserved) topic-specific Poisson variates.

A word’s topic-specific frequency, $\beta_{fk} \equiv \exp \mu_{fk}$, is directly parameterized in the model and is regularized across words (via hyperparameters ψ and γ^2) and across topics. A word’s exclusivity to a topic, $\phi_{f,k}$, is its usage rate relative to a set of comparison topics \mathcal{S} : $\phi_{f,k} = \beta_{f,k} / \sum_{j \in \mathcal{S}} \beta_{f,j}$. A topic’s siblings are a natural choice for a comparison set to see which words are overexpressed in the topic compared to a set of similar topics. While not directly modeled in HPC, the exclusivity parameters are also regularized by the τ_{fp}^2 , since if the child rates are forced to be similar then the $\phi_{f,k}$ will be pushed toward a baseline value of $1/|\mathcal{S}|$. We explore the regularization structure of the model empirically in Section 4.

Since both frequency and exclusivity are important factors in determining a word’s semantic content, a univariate measure of topical importance is a useful estimand for diverse tasks such as dimensionality reduction, feature selection, and content discovery. In constructing a composite measure, we do not want a high rank in one dimension to be able to compensate for a low rank in the other since frequency or exclusivity alone are not necessarily useful. We therefore adopt the harmonic mean to pull the “average” rank toward the lower score. For word f in topic k , we define the $FREX_{fk}$ score as the harmonic mean of the word’s rank in the distribution of $\phi_{.,k}$ and $\mu_{.,k}$:

$$FREX_{fk} = \left(\frac{w}{\text{ECDF}_{\phi_{.,k}}(\phi_{f,k})} + \frac{1-w}{\text{ECDF}_{\mu_{.,k}}(\mu_{f,k})} \right)^{-1}.$$

where w is the weight for exclusivity (which we set to 0.5 as a default) and $\text{ECDF}_{x_{.,k}}$ is the empirical CDF function applied to the values x over the first index.

3 Scalable inference via parallelized HMC sampler

We use a Gibbs sampler to obtain the posterior expectations of the unknown rate and membership parameters (and associated hyperparameters) given the observed data. Specifically, inference is conditioned on \mathbf{W} , a $D \times V$ matrix of word counts, \mathbf{I} , a $D \times K$ matrix of topic labels, \mathbf{l} , a D -vector of document lengths, and \mathcal{T} , a tree structure for the topics.

Creating a scalable inference method is critical since the space of latent variables grows linearly in the number of words and documents, with $K(D + V)$ total unknowns. Our model offers an advantage in that the posterior consists of two groups of parameters whose conditional posterior factors given the other. On one side, the conditional posterior of the rate and variance parameters $\{\mu_f, \tau_f^2\}_{f=1}^V$ factors by word given the membership parameters and the hyperparameters ψ, γ^2, ν and σ^2 . On the other, the conditional posterior of the topic affinity parameters $\{\xi_d\}_{d=1}^D$ factors by document given the hyperparameters η and Σ and the rate parameters $\{\mu_f\}_{f=1}^V$.

Conditional on the hyperparameters, therefore, we are left with two blocks of draws that can be broken into V or D independent threads. Using parallel computing software such as Message Passing Interface (MPI), the computation time for drawing the parameters in each block is only constrained by resources required for a single draw. The total runtime need not significantly increase with the addition of more documents or words as long as the number of available cores also increases.

Both of these conditional distributions are only known up to a constant and can be high dimensional if there are many topics, making direct sampling impossible and random walk Metropolis inefficient. We are able to obtain uncorrelated draws through the use of Hamiltonian Monte Carlo (HMC) (Neal, 2011), which leverages the posterior gradient and Hessian to find a distant point in the parameter space with high probability of acceptance. HMC works well for log densities that are unimodal and have relatively constant curvature. We give step-by-step instructions for our implementation of the algorithm in the Appendix.

After appropriate initialization, we follow a fixed Gibbs scan where the two blocks of latent variables are drawn in parallel from their conditional posteriors using HMC. We then draw the hyperparameters conditional on all the inputted latent variables.

3.1 Block Gibbs Sampler

To set up the block Gibbs sampling algorithm, we derive the relevant conditional posterior distributions and explain how we sample from each.

3.1.1 Updating tree parameters

In the first block, the conditional posterior of the tree parameters factors by word:

$$p(\{\boldsymbol{\mu}_f, \boldsymbol{\tau}_f^2\}_{f=1}^V | \mathbf{W}, \mathbf{I}, \mathbf{l}, \psi, \gamma^2, \nu, \sigma^2, \{\boldsymbol{\xi}_d\}_{d=1}^D, \mathcal{T}) \propto \prod_{f=1}^V \left\{ \prod_{d=1}^D p(w_{fd} | \mathbf{I}_d, l_d, \boldsymbol{\mu}_f, \boldsymbol{\xi}_d) \right\} \cdot p(\boldsymbol{\mu}_f, \boldsymbol{\tau}_f^2 | \psi, \gamma^2, \mathcal{T}, \nu, \sigma^2).$$

Given the conditional conjugacy of the variance parameters and their strong influence on the curvature of the rate parameter posterior, we sample the two groups conditional on each other to optimize HMC performance. Conditioning on the variance parameters, we can write the likelihood of the rate parameters as a Poisson regression where the documents are observations, the $\boldsymbol{\theta}_d(\mathbf{I}_d, \boldsymbol{\xi}_d)$ are the covariates, and the l_d serve as exposure weights.

The prior distribution of the rate parameters is a Gaussian graphical model, so *a priori* the log rates for each word are jointly Gaussian with mean $\psi \mathbf{1}$ and precision matrix $\boldsymbol{\Lambda}(\gamma^2, \boldsymbol{\tau}_f^2, \mathcal{T})$ which has non-zero entries only for topic pairs that have a direct parent-child relationship.² The log conditional posterior is:

$$\log p(\boldsymbol{\mu}_f | \mathbf{W}, \mathbf{I}, \mathbf{l}, \{\boldsymbol{\tau}_f^2\}_{f=1}^V, \psi, \gamma^2, \nu, \sigma^2, \{\boldsymbol{\xi}_d\}_{d=1}^D, \mathcal{T}) = - \sum_{d=1}^D l_d \boldsymbol{\theta}_d^T \boldsymbol{\beta}_f + \sum_{d=1}^D w_{fd} \log(\boldsymbol{\theta}_d^T \boldsymbol{\beta}_f) - \frac{1}{2} (\boldsymbol{\mu}_f - \psi \mathbf{1})^T \boldsymbol{\Lambda} (\boldsymbol{\mu}_f - \psi \mathbf{1}).$$

²In practice this precision matrix can be found easily as the negative Hessian of the log prior distribution.

We use HMC to sample from this unnormalized density. Note that the covariate matrix $\Theta_{D \times K}$ is very sparse in most cases, so we speed computation with a sparse matrix representation.

We know the conditional distribution of the variance parameters due to the conjugacy of the Inverse- χ^2 prior with the normal distribution of the log rates. Specifically, if $\mathcal{C}(\mathcal{T})$ is the set of child topics of topic k with cardinality J , then

$$\tau_{fk}^2 | \mu_f, \nu, \sigma^2, \mathcal{T} \sim \text{Inv-}\chi^2 \left(J + \nu, \frac{\nu\sigma^2 + \sum_{j \in \mathcal{C}} (\mu_{fj} - \mu_{fk})^2}{J + \nu} \right).$$

3.1.2 Updating topic affinity parameters

In the second block, the conditional posterior of the topic affinity vectors factors by document:

$$p(\{\xi_d\}_{d=1}^D | \mathbf{W}, \mathbf{I}, \mathbf{l}, \{\mu_f\}_{f=1}^V, \boldsymbol{\eta}, \boldsymbol{\Sigma}) \propto \prod_{d=1}^D \left\{ \prod_{f=1}^V p(w_{fd} | \mathbf{I}_d, l_d, \mu_f, \xi_d) \right\} \cdot p(\mathbf{I}_d | \xi_d) \cdot p(\xi_d | \boldsymbol{\eta}, \boldsymbol{\Sigma}).$$

We can again write the likelihood as a Poisson regression, now with the rates as covariates. The log conditional posterior for one document is:

$$\begin{aligned} \log p(\xi_d | \mathbf{W}, \mathbf{I}, \mathbf{l}, \{\mu_f\}_{f=1}^V, \boldsymbol{\eta}, \boldsymbol{\Sigma}) = & \\ & - l_d \sum_{f=1}^V \beta_f^T \boldsymbol{\theta}_d + \sum_{f=1}^V w_{fd} \log(\beta_f^T \boldsymbol{\theta}_d) - \sum_{k=1}^K \log(1 + e^{-\xi_{dk}}) \\ & - \sum_{k=1}^K (1 - I_{dk}) \xi_{dk} - \frac{1}{2} (\xi_d - \boldsymbol{\eta})^T \boldsymbol{\Sigma}^{-1} (\xi_d - \boldsymbol{\eta}). \end{aligned}$$

We use HMC to sample from this unnormalized density. Here the parameter vector $\boldsymbol{\theta}_d$ is sparse rather than the covariate matrix $\mathbf{B}_{V \times K}$. If we remove the entries of $\boldsymbol{\theta}_d$ and columns of \mathbf{B} pertaining to topics k where $I_{dk} = 0$, then we are left with a low dimensional regression where only the active topics are used as covariates, greatly simplifying computation.

3.1.3 Updating corpus-level parameters

We draw the hyperparameters after each iteration of the block update. We put flat priors on these unknowns so that we can learn their most likely values from the data. As a result, their conditional posteriors only depend on the latent variables they generate.

The log corpus-level rates $\mu_{f,0}$ for each word follow a Gaussian distribution with mean ψ and

variance γ^2 . The conditional distribution of these hyperparameters is available in closed form:

$$\begin{aligned} \psi|\gamma^2, \{\mu_{f,0}\}_{f=1}^V &\sim \mathcal{N}\left(\frac{1}{V} \sum_{f=1}^V \mu_{f,0}, \frac{\gamma^2}{V}\right), \\ \text{and } \gamma^2|\psi, \{\mu_{f,0}\}_{f=1}^V &\sim \text{Inv-}\chi^2\left(V, \frac{1}{V} \sum_{f=1}^V (\mu_{f,0} - \psi)^2\right). \end{aligned}$$

The discrimination parameters τ_{fk}^2 independently follow an identical Scaled Inverse- χ^2 with convolution parameter ν and scale parameter σ^2 , while their inverse follows a Gamma($\kappa_\tau = \frac{\nu}{2}, \lambda_\tau = \frac{2}{\nu\sigma^2}$) distribution. We use HMC to sample from this unnormalized density. Specifically,

$$\begin{aligned} \log p(\kappa_\tau, \lambda_\tau | \{\tau_f^2\}_{f=1}^V, \mathcal{T}) &= (\kappa_\tau - 1) \sum_{f=1}^V \sum_{k \in \mathcal{P}} \log (\tau_{fk}^2)^{-1} \\ &\quad - |\mathcal{P}|V\kappa_\tau \log \lambda_\tau - |\mathcal{P}|V \log \Gamma(\kappa_\tau) - \frac{1}{\lambda_\tau} \sum_{f=1}^V \sum_{k \in \mathcal{P}} (\tau_{fk}^2)^{-1}, \end{aligned}$$

where $\mathcal{P}(\mathcal{T})$ is the set of parent topics on the tree. Each draw of $(\kappa_\tau, \lambda_\tau)$ is then transformed back to the (ν, σ^2) scale.

The document-specific topic affinity parameters ξ_d follow a Multivariate Normal distribution with mean parameter η and a covariance matrix parameterized in terms of a scalar, $\Sigma = \lambda^2 \mathbf{I}_K$. The conditional distribution of these hyperparameters is available in closed form. For efficiency, we choose to put a flat prior on $\log \lambda^2$ rather than the original scale, which allows us to marginalize out η from the conditional posterior of λ^2 :

$$\begin{aligned} \lambda^2 | \{\xi_d\}_{d=1}^D &\sim \text{Inv-}\chi^2\left(DK - 1, \frac{\sum_d \sum_k (\xi_{dk} - \bar{\xi}_k)^2}{DK - 1}\right), \\ \text{and } \eta | \lambda^2, \{\xi_d\}_{d=1}^D &\sim \mathcal{N}\left(\bar{\xi}, \frac{\lambda^2}{D} \mathbf{I}_K\right). \end{aligned}$$

3.2 Estimation

As discussed in Section 2.3, our estimands are the topic-specific frequency and exclusivity of the words in the vocabulary, as well as the FREX score that averages each word's performance in these dimensions. We use posterior means to estimate frequency and exclusivity, computing these quantities at every iteration of the Gibbs sampler and averaging the draws after the burn-in period. For the FREX score, we applied the ECDF function to the frequency and exclusivity posterior expectations of all words in the vocabulary to estimate the true ECDF.

3.3 Inference for unlabeled documents

In order to classify unlabeled documents, we need to find the posterior predictive distribution of the membership vector $\mathbf{I}_{\tilde{d}}$ for a new document \tilde{d} . Inference is based on the new document's word counts $\mathbf{w}_{\tilde{d}}$ and the unknown parameters, which we hold constant at their posterior expectation. Unfortunately, the posterior predictive distribution of the topic affinities $\boldsymbol{\xi}_{\tilde{d}}$ is intractable without conditioning on the label vector since the labels control which topics contribute content. We therefore use a simpler model where the topic proportions depend only on the relative size of the affinity parameters:

$$\theta_{dk}^*(\boldsymbol{\xi}_d) \equiv \frac{e^{\xi_{dk}}}{\sum_{j=1}^K e^{\xi_{dj}}} \quad \text{and} \quad I_{dk} \sim \text{Bern}\left(\frac{1}{1 + \exp(-\xi_{dk})}\right).$$

The posterior predictive distribution of this simpler model factors into tractable components:

$$\begin{aligned} p^*(\mathbf{I}_{\tilde{d}}, \boldsymbol{\xi}_{\tilde{d}} | \mathbf{w}_{\tilde{d}}, \mathbf{W}, \mathbf{I}) &\approx p(\mathbf{I}_{\tilde{d}} | \boldsymbol{\xi}_{\tilde{d}}) p^*(\boldsymbol{\xi}_{\tilde{d}} | \{\hat{\boldsymbol{\mu}}_f\}_{f=1}^V, \hat{\boldsymbol{\eta}}, \hat{\boldsymbol{\Sigma}}, \mathbf{w}_{\tilde{d}}) \\ &\propto p(\mathbf{I}_{\tilde{d}} | \boldsymbol{\xi}_{\tilde{d}}) p^*(\mathbf{w}_{\tilde{d}} | \boldsymbol{\xi}_{\tilde{d}}, \{\hat{\boldsymbol{\mu}}_f\}_{f=1}^V) p(\boldsymbol{\xi}_{\tilde{d}} | \hat{\boldsymbol{\eta}}, \hat{\boldsymbol{\Sigma}}). \end{aligned}$$

It is then possible to find the most likely $\boldsymbol{\xi}_{\tilde{d}}^*$ based on the evidence from $\mathbf{w}_{\tilde{d}}$ alone.

4 Results

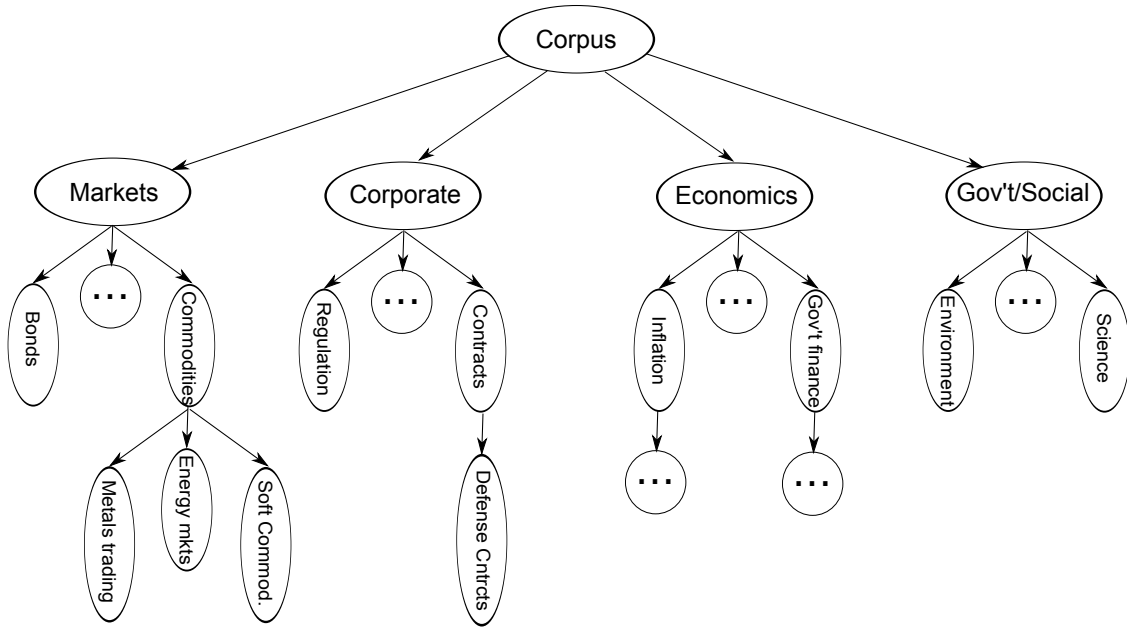
We analyze the fit of the HPC model to Reuters Corpus Volume I (RCV1), a large collection of newswire stories. First, we demonstrate how the variance parameters τ_{fp}^2 regularize the exclusivity with which words are expressed within topics. Second, we show that regularization of exclusivity has the greatest effect on infrequent words. Third, we explore the joint posterior of the topic-specific frequency and exclusivity of words as a summary of topical content, giving special attention to the upper right corner of the plot where words score highly in both dimensions. We compare words that score highly on the FREX metric to top words scored by frequency alone, the current practice in topic modeling. Finally, we compare the classification performance of HPC to baseline models.

4.1 The Reuters Corpus dataset

RCV1 is an archive of 806,791 newswire stories from a twelve-month period in 1996-1997.³ As described in [Lewis et al. \(2004\)](#), Reuters staffers assigned stories into any subset of 102 hierarchical topic categories. In the original data, assignment to any topic required automatic assignment to all ancestor nodes, but we removed these redundant ancestor labels since they do not allow our model to distinguish intentional assignments to high level categories from assignment to their offspring. In our modified annotations, the only documents we see in high level topics are those labeled to them and none of their children, which maps onto general content. We preprocessed document

³Available upon request from the National Institute of Standards and Technology (NIST), <http://trec.nist.gov/data/reuters/reuters.html>

Figure 2: Topic hierarchy of Reuters corpus



tokens with the Porter stemming algorithm (getting 300,166 unique stems) and chose the most frequent 3% of stems (10,421 unique stems, over 100 million total tokens) for the feature set.⁴

The Reuters topic hierarchy has three levels that divide the content into finer categories at each cut. At the first level, content is divided between four high level categories: three that focus on business and market news (Markets, Corporate/Industrial, and Economics) and one grab bag category that collects all remaining topics from politics to entertainment (Government/Social). The second level provides fine-grained divisions of these broad categories and contains the terminal nodes for most branches of the tree. For example, the Markets topic is split between equity, bond, money, and commodity markets at the second level. The third level offers further subcategories where needed for a small set of second level topics. For example, the Commodity Markets topic is divided between agricultural (soft), metal, and energy commodities. We present a graphical illustration of the Reuters topic hierarchy in Figure 2.

Many documents in the Reuters corpus are labeled to multiple topics, even after redundant ancestor memberships are removed. Overall, 32% of the documents are labeled to more than one node of the topic hierarchy. Fifteen percent of documents have very diverse content, being labeled to two or more of the main branches of the tree (Markets, Commerce, Economics, and Government/Social). Twenty-one percent of documents are labeled to multiple second-level categories on the same branch (for example, bond markets and equity markets in the Markets branch). Finally, 14% of documents are labeled to multiple children of the same second-level topic (for example, metals trading and energy markets in the commodity markets branch of Markets). Therefore, a completely general mixed membership model such as HPC is necessary to capture the labeling

⁴Including rarer features did not meaningfully change the results.

Table 2: Topic membership statistics

Topic code	Topic name	# docs	Any MM	CB L1 MM	CB L2 MM	CB L3 MM
CCAT	CORPORATE/INDUSTRIAL	2170	79.60%	79.60%	13.10%	0.80%
C11	STRATEGY/PLANS	24325	51.50	11.50	44.50	4.50
C12	LEGAL/JUDICIAL	11944	99.20	98.90	50.20	1.70
C13	REGULATION/POLICY	37410	85.90	55.60	61.40	4.50
C14	SHARE LISTINGS	7410	30.30	7.90	10.30	15.80
C15	PERFORMANCE	229	82.10	35.80	74.20	1.70
C151	ACCOUNTS/EARNINGS	81891	7.90	1.30	0.60	6.40
C152	COMMENT/FORECASTS	73092	18.90	4.80	1.60	13.50
C16	INSOLVENCY/LIQUIDITY	1920	66.70	31.50	54.60	3.60
C17	FUNDING/CAPITAL	4767	78.10	41.40	67.70	5.00
C171	SHARE CAPITAL	18313	44.60	3.20	1.70	41.50
C172	BONDS/DEBT ISSUES	11487	15.10	5.70	0.30	9.70
C173	LOANS/CREDITS	2636	24.70	8.50	3.60	15.60
C174	CREDIT RATINGS	5871	65.60	59.00	0.50	7.50
C18	OWNERSHIP CHANGES	30	76.70	23.30	76.70	3.30
C181	MERGERS/ACQUISITIONS	43374	34.40	6.50	4.80	26.90
C182	ASSET TRANSFERS	4671	28.30	4.70	5.70	21.00
C183	PRIVATISATIONS	7406	73.70	34.20	6.30	44.10
C21	PRODUCTION/SERVICES	25403	76.40	46.50	53.60	0.80
C22	NEW PRODUCTS/SERVICES	6119	55.00	15.30	49.10	0.40
C23	RESEARCH/DEVELOPMENT	2625	77.00	36.40	57.80	0.90
C24	CAPACITY/FACILITIES	32153	72.20	33.60	58.40	0.90
C31	MARKETS/MARKETING	29073	46.90	25.30	34.60	1.30
C311	DOMESTIC MARKETS	4299	80.60	73.70	9.50	18.70
C312	EXTERNAL MARKETS	6648	78.10	70.40	9.60	14.20
C313	MARKET SHARE	1115	39.70	10.30	5.10	27.80
C32	ADVERTISING/PROMOTION	2084	63.80	26.90	52.50	1.40
C33	CONTRACTS/ORDERS	14122	48.00	12.60	40.50	0.80
C331	DEFENCE CONTRACTS	1210	68.00	65.50	13.30	3.40
C34	MONOPOLIES/COMPETITION	4835	92.30	54.90	75.70	14.00
C41	MANAGEMENT	1083	75.60	52.10	59.90	2.00
C411	MANAGEMENT MOVES	10272	17.70	9.60	2.40	8.20
C42	LABOUR	11878	99.70	99.60	46.50	1.50
ECAT	ECONOMICS	621	90.50	90.50	9.70	1.40
E11	ECONOMIC PERFORMANCE	8568	43.00	24.20	29.10	5.10
E12	MONETARY/ECONOMIC	24918	81.70	75.40	17.90	13.70
E121	MONEY SUPPLY	2182	30.50	23.10	0.70	9.20
E13	INFLATION/PRICES	130	60.00	46.90	28.50	0.80
E131	CONSUMER PRICES	5659	24.70	15.60	6.00	12.00
E132	WHOLESALE PRICES	939	19.00	3.40	0.60	16.90
E14	CONSUMER FINANCE	428	73.80	43.20	61.00	1.60
E141	PERSONAL INCOME	376	75.00	63.80	9.60	22.30
E142	CONSUMER CREDIT	200	46.00	30.00	3.50	18.50
E143	RETAIL SALES	1206	27.50	19.70	2.40	10.20
E21	GOVERNMENT FINANCE	941	86.70	81.40	53.90	4.00
E211	EXPENDITURE/REVENUE	15768	78.20	72.40	16.10	13.80
E212	GOVERNMENT BORROWING	27405	32.70	29.60	2.70	4.50
E31	OUTPUT/CAPACITY	591	45.20	18.30	35.20	0.50
E311	INDUSTRIAL PRODUCTION	1701	17.70	9.80	3.10	9.30
E312	CAPACITY UTILIZATION	52	65.40	13.50	3.80	57.70
E313	INVENTORIES	111	26.10	10.80	0.00	16.20
E41	EMPLOYMENT/LABOUR	14899	100.00	100.00	49.40	2.20
E411	UNEMPLOYMENT	2136	92.00	90.60	10.40	12.00
E51	TRADE/RESERVES	4015	85.10	75.50	38.70	1.90
E511	BALANCE OF PAYMENTS	2933	63.80	43.70	8.20	25.70
E512	MERCHANDISE TRADE	12634	64.90	59.10	11.50	11.70
E513	RESERVES	2290	30.10	22.70	1.30	16.80
E61	HOUSING STARTS	391	51.70	47.80	13.80	0.80
E71	LEADING INDICATORS	5270	2.90	0.60	2.40	0.20

Key: MM = Mixed membership, CB Lx = Cross-branch MM at level x

Table 3: Topic membership statistics, con't

Topic code	Topic name	# docs	Any MM	CB L1 MM	CB L2 MM	CB L3 MM
GCAT	GOVERNMENT/SOCIAL	24546	2.50	2.50	0.50	0.10
G15	EUROPEAN COMMUNITY	1545	16.10	6.90	14.60	0.00
G151	EC INTERNAL MARKET	3307	98.00	87.20	10.60	94.30
G152	EC CORPORATE POLICY	2107	96.70	90.70	40.30	50.30
G153	EC AGRICULTURE POLICY	2360	96.10	94.20	31.40	27.70
G154	EC MONETARY/ECONOMIC	8404	98.20	93.00	11.50	43.90
G155	EC INSTITUTIONS	2124	70.80	42.00	24.30	54.00
G156	EC ENVIRONMENT ISSUES	260	75.00	57.70	28.80	50.80
G157	EC COMPETITION/SUBSIDY	2036	100.00	99.80	60.20	32.50
G158	EC EXTERNAL RELATIONS	4300	80.70	62.80	27.00	24.80
G159	EC GENERAL	40	47.50	17.50	35.00	2.50
GCRIM	CRIME, LAW ENFORCEMENT	32219	79.50	41.60	59.40	0.90
GDEF	DEFENCE	8842	93.70	17.20	84.40	0.50
GDIP	INTERNATIONAL RELATIONS	37739	73.70	20.50	60.70	0.90
GDIS	DISASTERS AND ACCIDENTS	8657	75.70	40.10	52.20	0.20
GENT	ARTS, CULTURE, ENTERTAINMENT	3801	68.80	29.20	49.60	0.50
GENV	ENVIRONMENT AND NATURAL WORLD	6261	90.20	51.50	72.30	2.50
GFAS	FASHION	313	76.40	45.70	41.50	1.90
GHEA	HEALTH	6030	81.90	56.10	65.00	1.20
GJOB	LABOUR ISSUES	17241	99.60	99.40	44.60	3.30
GMIL	MILLENNIUM ISSUES	5	100.00	100.00	40.00	0.00
GOBIT	OBITUARIES	844	99.40	15.30	99.40	0.00
GODD	HUMAN INTEREST	2802	60.70	9.70	55.20	0.10
GPOL	DOMESTIC POLITICS	56878	79.60	29.70	63.00	1.80
GPRO	BIOGRAPHIES, PERSONALITIES, PEOPLE	5498	87.50	10.00	84.70	0.10
GREL	RELIGION	2849	86.10	6.60	84.30	0.10
GSCI	SCIENCE AND TECHNOLOGY	2410	55.20	22.20	45.10	0.30
GSPO	SPORTS	35317	1.30	0.60	0.90	0.00
GTOUR	TRAVEL AND TOURISM	680	89.60	69.70	34.70	3.40
GVIO	WAR, CIVIL WAR	32615	67.30	10.10	64.60	0.10
GVOTE	ELECTIONS	11532	100.00	13.30	100.00	1.30
GWEA	WEATHER	3878	73.90	46.80	46.40	0.10
GWELF	WELFARE, SOCIAL SERVICES	1869	95.40	75.50	74.10	3.40
MCAT	MARKETS	894	81.10	81.10	14.50	2.20
M11	EQUITY MARKETS	48700	16.30	12.30	3.90	2.90
M12	BOND MARKETS	26036	21.30	15.60	5.20	3.50
M13	MONEY MARKETS	447	65.80	51.90	23.30	1.60
M131	INTERBANK MARKETS	28185	15.10	9.40	0.70	6.40
M132	FOREX MARKETS	26752	36.90	24.70	3.10	16.10
M14	COMMODITY MARKETS	4732	18.00	16.70	2.30	0.10
M141	SOFT COMMODITIES	47708	24.10	22.80	5.50	2.00
M142	METALS TRADING	12136	34.70	19.30	4.10	16.10
M143	ENERGY MARKETS	21957	21.10	18.40	4.80	2.90

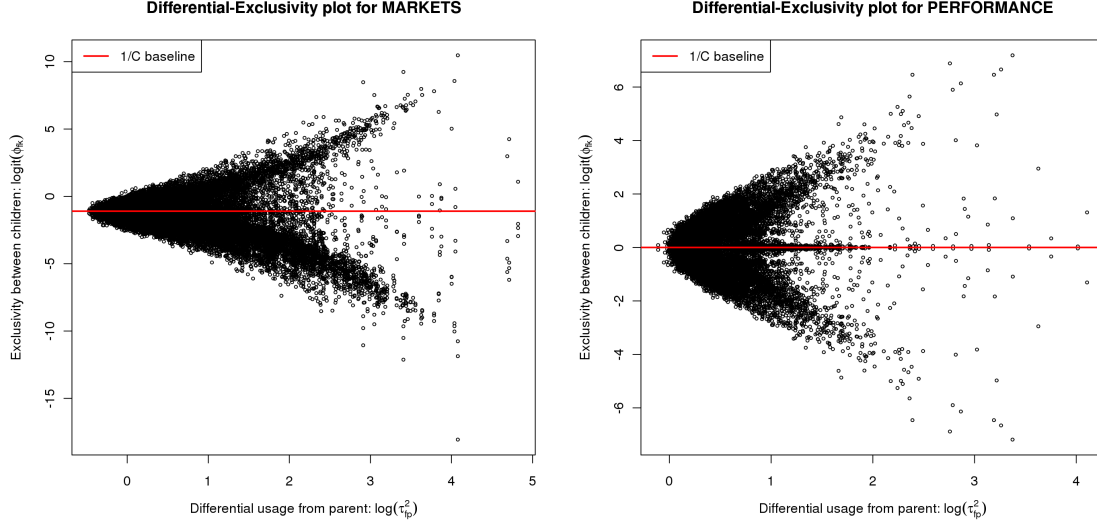
Key: MM = Mixed membership, CB Lx = Cross-branch MM at level x

patterns of the corpus. A full breakdown of membership statistics by topic is presented in Tables 2 and 3.

4.2 How the differential usage parameters regulate topic exclusivity

A word can only be exclusive to a topic if its expression across the sibling topics is allowed to diverge from the parent rate. Therefore, we would only expect words with high differential usage parameters τ_{fp}^2 at the parent level to be candidates for highly exclusive expression ϕ_{fk} in any child topic k . Words with child topic rates that cannot vary greatly from the parent should have

Figure 3: Exclusivity as a function of differential usage parameters



nearly equal expression in each child k , meaning $\phi_{fk} \approx \frac{1}{C}$ for a branch with C child topics. An important consequence is that, although the ϕ_{fk} are not directly modeled in HPC, their distribution is regularized by learning a prior distribution on the τ_{fp}^2 .

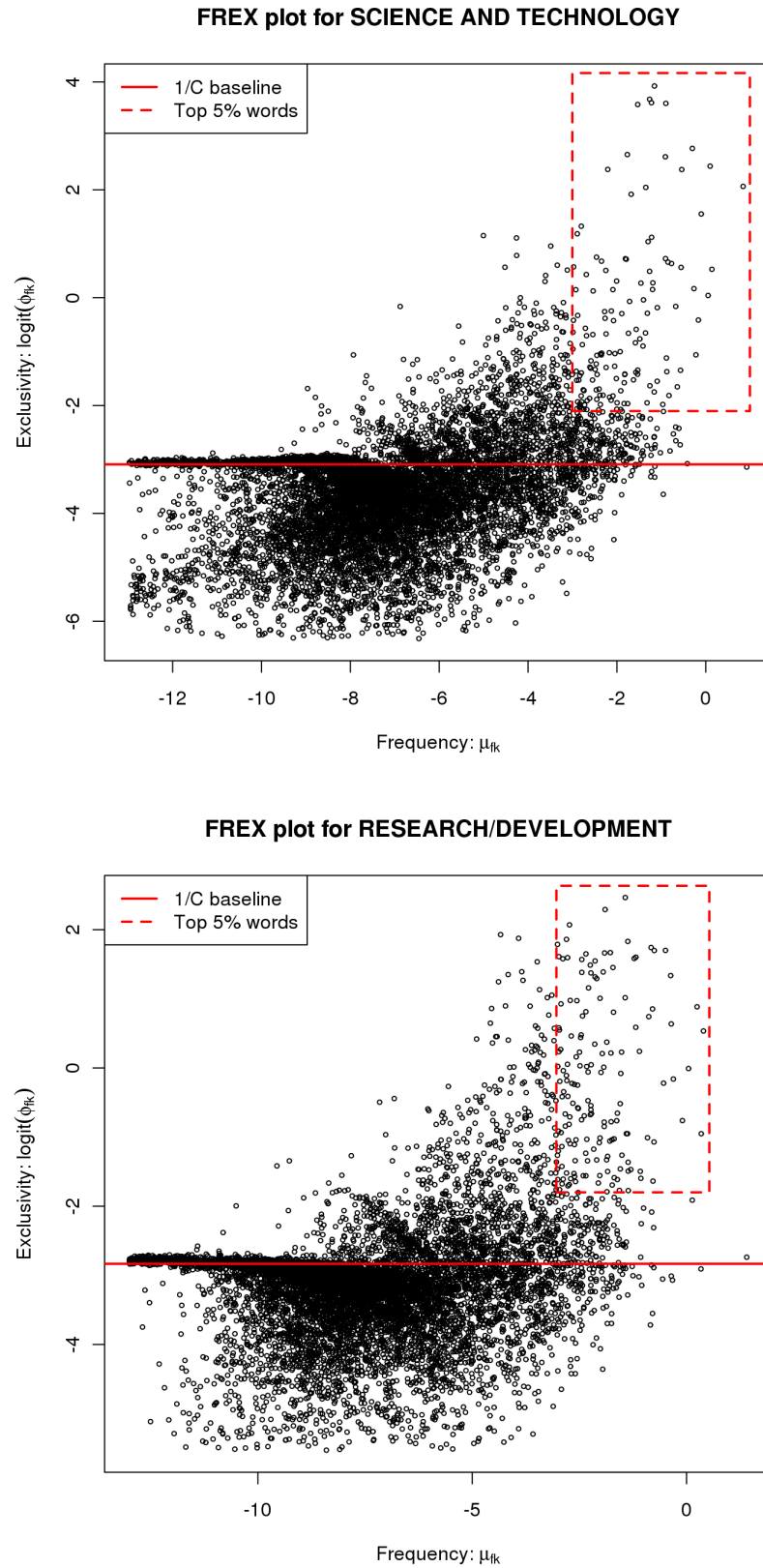
This tight relation can be seen in the HPC fit. Figure 3 shows the joint posterior expectation of the differential usage parameters in a parent topic and exclusivity parameters across the child topics. Specifically, the left panel compares the rate variance of the children of Markets from their parent to exclusivity between the child topics; the right panel does the same with the two children of Performance, a second-level topic under the Corporate category. The plots have similar patterns. For low levels of differential expression, the exclusivity parameters are clustered around the baseline value, $\frac{1}{C}$. At high levels of child rate variance, words gain the ability to approach exclusive expression in a single topic.

4.3 How frequency modulates regularization of exclusivity

One of the most appealing aspects of regularization in generative models is that it acts most strongly on the parameters for which we have the least information. In the case of the exclusivity parameters in HPC we have the most data for frequent words, so for a given topic the words with low rates should be least able to escape regularization of their exclusivity parameters by our shrinkage prior on the parent's τ_{fp}^2 .

Figure 4 shows for two topics the joint posterior expectation of each word's frequency in that topic and its exclusivity compared to sibling topics (the FREX plot). The left panel features the Science and Technology topic, a child in the grab bag Government/Social branch, and the right panel features the Research/Development topic, a child in the Corporate branch. The overall shape

Figure 4: Frequency-Exclusivity (FREX) plots



of the joint posterior is very similar for both topics. On the left side of the plots, the exclusivity of rare words is unable to significantly exceed the $\frac{1}{C}$ baseline. This is because the model does not have much evidence to estimate usage in the topic, so the estimated rate is shrunk heavily toward the parent rate. However, we see that it is possible for rare words to be underexpressed in a topic, which happens if they are frequent and overexpressed in a sibling topic. Even though their rates are similar to the parent in this topic, sibling topics may have a much higher rate and account for most appearances of the word in the comparison group.

4.4 Frequency and Exclusivity as a two dimensional summary of semantic content

Words in the upper right of the FREX plot—those that are both frequent and highly exclusive—are of greatest interest. These are the most common words in the corpus that are also likely to have been generated from the topic of interest (rather than similar topics). We show words in the upper 5% quantiles in both dimensions for our example topics in Figure 5. These high-scoring words can help to clarify content even for labeled topics. In the Science and Technology topic, we see almost all terms are specific to the American and Russian space programs. Similarly, in the Research/Technology topic, almost all terms relate to clinical trials in medicine or to agricultural research.

We also compute the Frequency-Exclusivity (FREX) score for each word-topic pair, a univariate summary of topical content that averages performance in both dimensions. In Table 4 we compare the top FREX words in three topics to a ranking based on frequency alone, which is the current practice in topic modeling. For context, we also show the immediate neighbors of each topic in the tree. The topic being examined is in bolded red, while the borders of the comparison set are solid. The Defense Contracts topic is a special case since it is an only child. In these cases, we use a comparison to the parent topic to calculate exclusivity.

By incorporating exclusivity information, FREX-ranked lists include fewer words that are used similarly everywhere (such as *said* and *would*) and fewer words that are used similarly in a set of related topics (such as *price* and *market* in the Markets branch). One can understand this result by comparing the rankings for known stop words from the SMART list to other words. In Figure 6, we show the maximum ECDF ranking for each word across topics in the distribution of frequency (left panel) and exclusivity (right panel) estimates. One can see that while stop words are more likely to be in the extreme quantiles of frequency, very few of them are among the most exclusive words. This prevents general and context-specific stop words from ranking highly in a FREX-based index.

4.5 Classification performance

We compare the classification performance of HPC with SVM and L2-regularized logistic regression (Genkin et al., 2007; Rubin et al., 2012; Ghamrawi and McCallum, 2005). All methods were trained on a random sample of 15% of the documents using the 3% most frequent words in the corpus as features. These fits were used to predict memberships in the withheld documents, an

Upper 5% of FREX plot for SCIENCE AND TECHNOLOGY



Table 4: Comparison of High FREX words (both frequent and exclusive) to most frequent words
(featured topic name bold red; comparison set in solid ovals)

	High FREX	Most frequent
Metals Trading	copper	said
	aluminium	gold
	metal	price
	gold	copper
	zinc	market
	ounc	metal
	silver	trader
	palladium	tonn
	comex	trade
	platinum	close
	bullion	ounc
	preciou	aluminium
	nickel	london
	mine	dealer
Environment	greenpeac	said
	environment	would
	pollut	environment
	wast	year
	emiss	state
	reactor	nuclear
	forest	million
	speci	greenpeac
	environ	world
	eleph	water
	spill	group
	wildlif	govern
	energi	nation
	nuclear	environ
Defense Contracts	fighter	said
	defenc	contract
	missil	million
	forc	system
	defens	forc
	eurofight	defenc
	armi	would
	helicopt	aircraft
	lockhe	compani
	czech	deal
	martin	fighter
	militari	govern
	navi	unit
	mcdonnel	lockhe

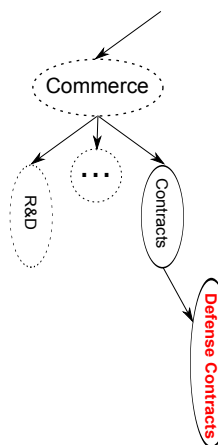
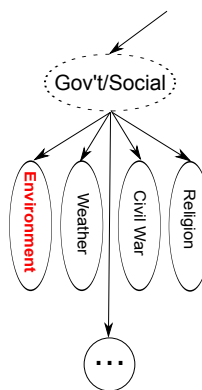
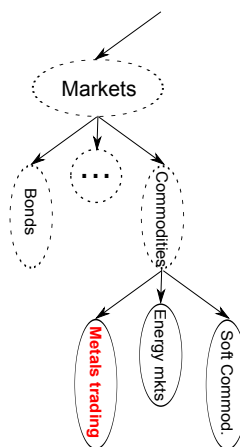
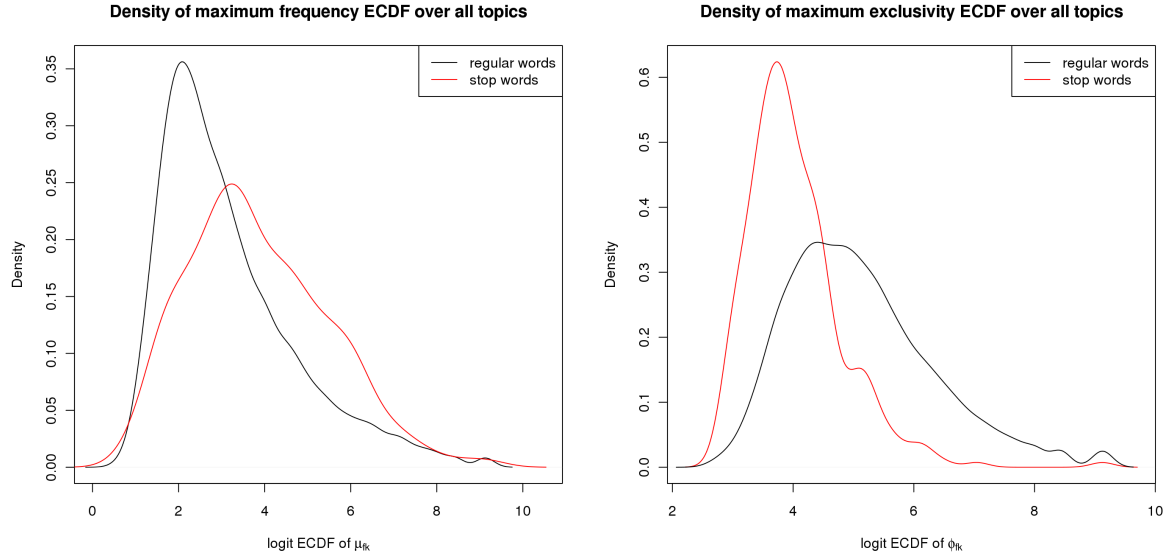


Figure 6: Comparison of FREX score components for SMART stop words vs. regular words



experiment we repeated ten times with a new random sample as a training set. Table 5 shows the results of our experiment, using both micro averages (every document weighted equally) and macro averages (every topic weighted equally). While HPC does not dominate other methods, on average its performance does not deviate significantly from traditional classification algorithms.

HPC is not designed for optimizing predictive accuracy out-of-sample, rather it is designed to maximize interpretability of the label-specific summaries, in terms of words that are both frequent and exclusive. These results offer a quantitative illustration of the classical trade-off between predictive and explanatory power of statistical models (Breiman, 2001).

Table 5: Classification performance for ten-fold cross-validation

	SVM	L2-reg Logit	HPC
Micro-ave Precision	0.711 (0.002)	0.195 (0.031)	0.695 (0.007)
Micro-ave Recall	0.706 (0.001)	0.768 (0.013)	0.589 (0.008)
Macro-ave Precision	0.563 (0.002)	0.481 (0.025)	0.505 (0.094)
Macro-ave Recall	0.551 (0.006)	0.600 (0.007)	0.524 (0.093)

Standard deviation of performance over ten folds in parenthesis.

5 Discussion

Our thesis is that one needs to know how words are used differentially across topics as well as within them in order to understand topical content; we refer to these dimensions of content as word exclusivity and frequency. Topical summaries that focus on word frequency alone are often dominated by stop words or other terms used similarly across many topics. Exclusivity and frequency can be visualized graphically as a latent space or combined into an index such as the FREX score to obtain a univariate measure of the topical content for words in each topic.

Naive estimates of exclusivity will be biased toward rare words due to sensitivity to small differences in estimated use across topics. Existing topic models such as LDA cannot regularize differential use due to topic normalization of usage rates; its symmetric Dirichlet prior on topic distributions regularizes within, not between, topic usage. While topic-regularized models can capture many important facets of word usage, they are not optimal for the estimands used in our analysis of topical content.

HPC breaks from standard topic models by modeling topic-specific word counts as unnormalized count variates whose rates can be regularized both within and across topics to compute word frequency and exclusivity. It was specifically designed to produce stable exclusivity estimates in human-annotated corpora by smoothing differential word usage according to a semantically intelligent distance metric: proximity on a known hierarchy. This supervised setting is an ideal test case for our framework and will be applicable to many high value corpora such as the *ACM library*, *IMS* publications, the *New York Times* and *Reuters*, which all have professional editors and authors and provide multiple annotations to a hierarchy of labels for each document.

HPC offers a complex challenge for full Bayesian inference. To offer a flexible framework for regularization, it breaks from the simple Dirichlet-Multinomial conjugacy of traditional models. Specifically, HPC uses Poisson likelihoods whose rates are smoothed across a known topic hierarchy with a Gaussian diffusion and a novel mixed membership model where document label and topic membership parameters share a Gaussian prior. The membership model is the first to create an explicit link between the distribution of topic labels in a document and of the words that appear in a document and allow for multiple labels. However, the resulting inference is challenging since, conditional on word usage rates, the posterior of the membership parameters involves Poisson and Bernoulli likelihoods of differing dimensions constrained by a Gaussian prior.

We offer two methodological innovations to make inference tractable. First, we design our model with parameters that divide cleanly into two blocks (the tree and document parameters) whose members are conditionally independent given the other block, allowing for parallelized, scalable inference. However, these factorized distributions cannot be normalized analytically and are the same dimension as the number of topics (102 in the case of *Reuters*). We therefore implement a Hamiltonian Monte Carlo conditional sampler that mixes efficiently through high dimensional spaces by leveraging the posterior gradient and Hessian information. This allows HPC to scale to large and complex topic hierarchies that would be intractable for Random Walk Metropolis samplers.

One unresolved bottleneck in our inference strategy is that the MCMC sampler mixes slowly

through the hyperparameter space of the documents—the η and λ^2 parameters that control the mean and sparsity of topic memberships and labels. This is due to a large fraction of missing information in our augmentation strategy (Meng and Rubin, 1991). Conditional on all the documents’ topic affinity parameters $\{\xi_d\}_{d=1}^D$, these hyperparameters index a normal distribution with D observations; marginally, however, we have much less information about the exact loading of each topic onto each document. While we have been exploring more efficient data augmentation strategies such as Parameter Expansion (Liu and Wu, 1999), we have not found a workable alternative to augmenting the posterior with the entire set of $\{\xi_d\}_{d=1}^D$ parameters.

5.1 Concluding remarks

While HPC was developed for the specific case of hierarchically labeled document collections, this framework can be readily extended to other types of document corpora. For labeled corpora where no hierarchical structure on the topics is available, one can use a flat hierarchy to model differential use. For document corpora where no labeled examples are available, a simple word rate model with a flat hierarchy and dense topic membership structure could be employed to get more informative summaries of inferred topics. In either case, the word rate framework could be combined with non-parameteric Bayesian models that infer hierarchical structure on the topics (Adams et al., 2010). We expect modeling approaches based on rates will play an important role in future work on text summarization.

The HPC model can also be leveraged to semi-automate the construction of topic ontologies targeted to specific domains, for instance, when fit to comprehensive human-annotated corpora such as *Wikipedia*, *The New York Times*, *Encyclopedia Britannica*, or databases such as *JSTOR* and the *ACM repository*. By learning a probabilistic representation of high quality topics, HPC output can be used as a gold standard to aid and evaluate other learning methods. Targeted ontologies have been a key factor in monitoring scientific progress in biology (Ashburner et al., 2000; Kanehisa and Goto, 2000). A hierarchical ontology of topics would lead to new metrics for measuring progress in text analysis. It would enable an evaluation of the semantic content of any collection of inferred topics, thus finally allowing for a *quantitative comparison* among the output of topic models. Current evaluations are qualitative, anecdotal and unsatisfactory; for instance, authors argue that lists of most frequent words describing an arbitrary selection of topics inferred by a new model make sense intuitively, or that they are better than lists obtained with other models.

In addition to model evaluation, a news-specific ontology could be used as prior to inform the analysis of unstructured text, including Twitter feeds, Facebook wall posts, and blogs. Unsupervised topic models infer a latent topic space that may be oriented around unhelpful axes, such as authorship or geography. Using a human-created ontology as a prior could ensure that a useful topic space is discovered without being so dogmatic as to assume that unlabeled documents have the same latent structure as labeled examples.

References

- R. P. Adams, Z. Ghahramani, and M. I. Jordan. Tree-structured stick breaking for hierarchical data. In J. Shawe-Taylor, R. Zemel, J. Lafferty, and C. Williams, editors, *Advances in Neural Information Processing (NIPS) 23*, 2010.
- E. M. Airoldi, W. W. Cohen, and S. E. Feinberg. Bayesian methods for frequent terms in text: Models of contagion and the delta-square statistic. CSNA and INTERFACE Annual Meetings, 2005.
- E. M. Airoldi, A. G. Anderson, S. E. Fienberg, and K. K. Skinner. Who wrote Ronald Reagan’s radio addresses? *Bayesian Analysis*, 1(2):289–320, 2006.
- E. M. Airoldi, S. E. Fienberg, and K. K. Skinner. Whose ideas? Whose words? Authorship of the Ronald Reagan radio addresses. *Political Science & Politics*, 40:501–506, 2007.
- E. M. Airoldi, D. M. Blei, S.E. Fienberg, and E.P. Xing. Mixed-membership stochastic blockmodels. *Journal of Machine Learning Research*, 9:1981–2014, 2008.
- E. M. Airoldi, E. A. Erosheva, S. E. Fienberg, C. J. Joutard, T. M. Love, and S. Shringarpure. Reconceptualizing the classification of pnas articles. *PNAS*, 107, 2010.
- M. Ashburner, C. A. Ball, J. A. Blake, D. Botstein, H. Butler, J. M. Cherry, A. P. Davis, K. Dolinski, S. S. Dwight, J. T. Eppig, M. A. Harris, D. P. Hill, L. Issel-Tarver, A. Kasarskis, S. Lewis, J. C. Matese, J. E. Richardson, M. Ringwald, G. M. Rubinand, and G. Sherlock. Gene ontology: Tool for the unification of biology. The gene ontology consortium. *Nature Genetics*, 25(1): 25–29, 2000.
- Jonathan Bischof and Edoardo Airoldi. Summarizing topical content with word frequency and exclusivity. ICML, 2012.
- D. Blei. Introduction to probabilistic topic models. *Communications of the ACM*, 2012. In press.
- D. Blei, T. Griffiths, M. Jordan, and J. Tenenbaum. Hierarchical topic models and the nested Chinese restaurant process. NIPS, 2003a.
- David Blei and John McAuliffe. Supervised topic models. volume 21. Neural Information Processing Systems, 2007.
- David Blei, Andrew Ng, and Michael Jordan. Latent dirichlet allocation. *Journal of Machine Learning Research*, 2003b.
- L. Breiman. Statistical modeling: The two cultures. *Statistical Science*, 16(3):199–231, 2001.
- John Canny. GAP: A Factor Model for Discrete Data. SIGIR, 2004.
- Jonathan Chang, Jordan Boyd-Graber, Sean Gerrish, Chong Wang, and David Blei. Reading tea leaves: How humans interpret topic models. Neural Information Processing Systems, 2009.

- Jacob Eisenstein, Amr Ahmed, and Eric P. Xing. Sparse Additive Generative Models of Text. ICML, 2011.
- Alexander Genkin, David D. Lewis, and David Madigan. Large-scale bayesian logistic regression for text categorization. *Technometrics*, 49, 2007.
- Nadia Ghamrawi and Andrew McCallum. Collective multi-label classification. Fourteenth Conference on Information and Knowledge Management (CIKM), 2005.
- Justin Grimmer and Gary King. General purpose computer-assisted clustering and conceptualization. *PNAS*, 2011.
- H. Hotelling. Relations between two sets of variants. *Biometrika*, 28:321–377, 1936.
- Yuening Hu, Jordan Boyd-Graber, and Brianna Satinoff. Interactive Topic Modeling. Association for Computational Linguistics, 2011.
- I. T. Jolliffe. *Principal Component Analysis*. Springer-Verlag, 1986.
- M. Kanehisa and S. Goto. KEGG: Kyoto encyclopedia of genes and genomes. *Nucleic Acids Research*, 28(1):27–30, 2000.
- David D. Lewis, Yiming Yang, Tony G. Rose, and Fan Li. RCV1: A New Benchmark Collection for Text Categorization Research. *Journal of Machine Learning Research*, 5:361–397, 2004.
- Jun S. Liu and Ying Nian Wu. Parameter expansion for data augmentation. *Journal of the American Statistical Association*, 94:1264–1274, 1999.
- Andrew McCallum, Ronald Rosenfeld, Tom Mitchell, and Andrew Ng. Improving text classification by shrinkage in a hierarchy of classes. International Conference on Machine Learning, 1998.
- Geoffrey McLachlan and David Peel. *Finite Mixture Models*. Wiley, 2000.
- Xiao-Li Meng and Donald Rubin. Using em to obtain asymptotic variance-covariance matrices: The sem algorithm. *Journal of the American Statistical Association*, 86:899–909, 1991.
- David Mimno, Wei Li, and Andrew McCallum. Mixtures of hierarchical topics with pachinko allocation. ICML, 2007.
- David Mimno, Hanna Wallach, Edmund Talley, Miriam Leenders, and Andrew McCallum. Optimizing Semantic Coherence in Topic Models. EMNLP, 2011.
- Burt Monroe, Michael Colaresi, and Kevin Quinn. Fightin’ words: Lexical feature selection and evaluation for identifying the content of political conflict. *Political Analysis*, 16:372–403, 2008.
- F. Mosteller and D.L. Wallace. *Applied Bayesian and Classical Inference: The Case of “The Federalist” Papers*. Springer-Verlag, 1984.

- Radford Neal. MCMC using Hamiltonian dynamics. In Steve Brooks, Andrew Gelman, Galin L. Jones, and Xiao-Li Meng, editors, *Handbook of Markov Chain Monte Carlo*. Chapman & Hall / CRC Press, 2011.
- Adler Perotte, Nicholas Bartlett, Noemie Elhadad, and Frank Wood. Hierarchically Supervised Latent Dirichlet Allocation. NIPS, 2012.
- Daniel Ramage, David Hall, Ramesh Nallapati, and Christopher D. Manning. Labeled LDA: A supervised topic model for credit attribution in multi-labeled corpora. EMNLP, 2009.
- T. Rubin, A. Chambers, P. Smyth, and M. Steyvers. Statistical topic models for multi-label document classification. *Machine Learning*, 88, 2012.
- Kyung-Ah Sohn and Eric P. Xing. A hierarchical dirichlet process mixture model for haplotype reconstruction from multi-population data. *Annals of Applied Statistics*, 3:791–821, 2009.
- Hanna Wallach, David Mimno, and Andrew McCallum. Rethinking LDA: Why Priors Matter. NIPS, 2009.
- Jun Zhu and Eric P. Xing. Sparse Topical Coding. UAI, 2011.

A Appendix: Implementing the parallelized HMC sampler

A.1 Hamiltonian Monte Carlo conditional updates

Hamiltonian Monte Carlo (HMC) is the key tool that makes high-dimensional, non-conjugate updates tractable for our Gibbs sampler. It works well for log densities that are unimodal and have relatively constant curvature. We outline our customized implementation of the algorithm here; a general introduction can be found in [Neal \(2011\)](#).

HMC is a version of the Metropolis-Hastings algorithm that replaces the common Multivariate Normal proposal distribution with a distribution based on Hamiltonian dynamics. It can be used to make joint proposals on the entire parameter space or, as in this paper, to make proposals along the conditional posteriors as part of a Gibbs scan. While it requires closed form calculation of the posterior gradient and curvature to perform well, the algorithm can produce uncorrelated or negatively correlated draws from the target distribution that are almost always accepted.

A consequence of classical mechanics, Hamiltonian’s equations can be used to model the movement of a particle along a frictionless surface. The total energy of the particle is the sum of its potential energy (the height of the surface relative to the minimum at the current position) and its kinetic energy (the amount of work needed to accelerate the particle from rest to its current velocity). Since energy is preserved in a closed system, the particle can only convert potential energy to kinetic (or vice versa) as it moves along the surface.

Imagine a ball placed high on the side of the parabola $f(q) = q^2$ at position $q = -2$. Starting out, it will have no kinetic energy but significant potential energy due to its position. As it rolls down the parabola toward zero, it speeds up (gaining kinetic energy), but loses potential energy to compensate as it moves to a lower position. At the bottom of the parabola the ball has only kinetic energy, which it then translates back into potential energy by rolling up the other side until its kinetic energy is exhausted. It will then roll back down the side it just climbed, completely reversing its trajectory until it returns to its original position.

HMC uses Hamiltonian dynamics as a method to find a distant point in the parameter space with high probability of acceptance. Suppose we want to produce samples from $f(\mathbf{q})$, a possibly unnormalized density. Since we want high probability regions to have the least potential energy, we parameterize the surface the particle moves along as $U(\mathbf{q}) = -\log f(\mathbf{q})$, which is the height of the surface and the potential energy of the particle at any position \mathbf{q} . The total energy of the particle, $H(\mathbf{p}, \mathbf{q})$, is the sum of its kinetic energy, $K(\mathbf{p})$, and its potential energy, $U(\mathbf{q})$, where \mathbf{p} is its momentum along each coordinate. After drawing an initial momentum for the particle (typically chosen as $\mathbf{p} \sim \mathcal{N}(\mathbf{0}, \mathbf{M})$, where \mathbf{M} is called the *mass matrix*), we allow the system to evolve for a period of time—not so little that there is negligible absolute movement, but not so much that the particle has time to roll back to where it started.

HMC will not generate good proposals if the particle is not given enough momentum in each direction to efficiently explore the parameter space in a fixed window of time. The higher the curvature of the surface, the more energy the particle needs to move to a distant point. Therefore the performance of the algorithm depends on having a good estimate of the posterior curvature

$\hat{H}(\mathbf{q})$ and drawing $\mathbf{p} \sim \mathcal{N}(\mathbf{0}, -\hat{H}(\mathbf{q}))$. If the estimated curvature is accurate and relatively constant across the parameter space, the particle will have high initial momentum along directions where the posterior is concentrated and less along those where the posterior is more diffuse.

Unless the (conditional) posterior is very well behaved, the Hessian should be calculated at the log-posterior mode to ensure positive definiteness. Maximization is generally an expensive operation, however, so it is not feasible to update the Hessian every iteration of the sampler. In contrast, the log-prior curvature is very easy to calculate and well behaved everywhere. This led us to develop the *scheduled conditional HMC sampler* (SCHMC), an algorithm for nonconjugate Gibbs draws that updates the log-prior curvature at every iteration but only updates the log-likelihood curvature in a strategically chosen subset of iterations. We use this algorithm for all non-conjugate conditional draws in our Gibbs sampler.

Specifically, suppose we want to draw from the conditional distribution $p(\boldsymbol{\theta}|\boldsymbol{\psi}_t, \mathbf{y}) \propto p(\mathbf{y}|\boldsymbol{\theta}, \boldsymbol{\psi}_t)p(\boldsymbol{\theta}|\boldsymbol{\psi}_t)$ in each Gibbs scan, where $\boldsymbol{\psi}$ is a vector of the remaining parameters and \mathbf{y} is the observed data. Let \mathcal{S} be the set of full Gibbs scans in which the log-likelihood Hessian information is updated (which always includes the first). For Gibbs scan $i \in \mathcal{S}$, we first calculate the conditional posterior mode and evaluate both the Hessian of the log-likelihood, $\log p(\mathbf{y}|\boldsymbol{\theta}, \boldsymbol{\psi}_t)$, and of the log-prior, $\log p(\boldsymbol{\theta}|\boldsymbol{\psi}_t)$, at that mode, adding them together to get the log-posterior Hessian. We then get a conditional posterior draw with HMC using the negative Hessian as our mass matrix. For Gibbs scan $i \notin \mathcal{S}$, we evaluate the log-prior Hessian at the current location and add it our last evaluation of the log-likelihood Hessian to get the log-posterior Hessian. We then proceed as before. The SCHMC procedure is described in step-by-step detail in Algorithm 1.

A.2 SCHMC implementation details for HPC model

In the previous section we described our general procedure for obtaining samples from unnormalized conditional posteriors, the SCHMC algorithm. In this section, we provide the gradient and Hessian calculations necessary to implement this procedure for the unnormalized conditional densities in the HPC model, as well as strategies to obtain the maximum of each conditional posterior.

A.2.1 Conditional posterior of the rate parameters

The log conditional posterior of the rate parameters for one word is:

$$\begin{aligned} \log p(\boldsymbol{\mu}_f | \mathbf{W}, \mathbf{I}, \mathbf{l}, \{\boldsymbol{\tau}_f^2\}_{f=1}^V, \boldsymbol{\psi}, \gamma^2, \nu, \sigma^2, \{\boldsymbol{\xi}_d\}_{d=1}^D, \mathcal{T}) \\ = \sum_{d=1}^D \log \text{Pois}(w_{fd} | l_d \boldsymbol{\theta}_d^T \boldsymbol{\beta}_f) + \log \mathcal{N}(\boldsymbol{\mu}_f | \boldsymbol{\psi} \mathbf{1}, \boldsymbol{\Lambda}(\gamma^2, \boldsymbol{\tau}_f^2, \mathcal{T})) \\ = - \sum_{d=1}^D l_d \boldsymbol{\theta}_d^T \boldsymbol{\beta}_f + \sum_{d=1}^D w_{fd} \log(\boldsymbol{\theta}_d^T \boldsymbol{\beta}_f) - \frac{1}{2} (\boldsymbol{\mu}_f - \boldsymbol{\psi} \mathbf{1})^T \boldsymbol{\Lambda} (\boldsymbol{\mu}_f - \boldsymbol{\psi} \mathbf{1}). \end{aligned}$$

Algorithm 1: Scheduled conditional HMC sampler for iteration i

input : θ_{t-1}, ψ_t (current value of other parameters), \mathbf{y} (observed data), L (number of leapfrog steps), ϵ (stepsize), and \mathcal{S} (set of full Gibbs scans in which the likelihood Hessian is updated)

output: θ_t

$\theta_0^* \leftarrow \theta_{t-1};$

/* Update conditional likelihood Hessian if iteration in schedule */
if $i \in \mathcal{S}$ **then**
 $\hat{\theta} \leftarrow \arg \max_{\theta} \{\log p(\mathbf{y}|\theta, \psi_t) + \log p(\theta|\psi_t)\};$
 $\hat{H}_l(\theta) \leftarrow \frac{\partial^2}{\partial \theta \partial \theta^T} [\log p(\mathbf{y}|\hat{\theta}, \psi_t)]|_{\theta=\hat{\theta}};$
end

/* Calculate prior Hessian and set up mass matrix */
 $\hat{H}_p(\theta) \leftarrow \frac{\partial^2}{\partial \theta \partial \theta^T} [\log p(\theta|\psi_t)]|_{\theta=\theta_0^*};$
 $\hat{H}(\theta) \leftarrow \hat{H}_l(\theta) + \hat{H}_p(\theta);$
 $M \leftarrow -\hat{H}(\theta);$

/* Draw initial momentum */
Draw $\mathbf{p}_0^* \sim \mathcal{N}(\mathbf{0}, M);$

/* Leapfrog steps to get HMC proposal */
for $l \leftarrow 1$ **to** L **do**
 $\mathbf{g}_1 \leftarrow -\frac{\partial}{\partial \theta} [\log p(\theta|\psi_t, \mathbf{y})]|_{\theta=\theta_{l-1}^*};$
 $\mathbf{p}_{l,1}^* \leftarrow \mathbf{p}_{l-1}^* - \frac{\epsilon}{2} \mathbf{g}_1;$
 $\theta_l^* \leftarrow \theta_{l-1}^* + \epsilon(M^{-1})^T \mathbf{p}_{l,1}^*;$
 $\mathbf{g}_2 \leftarrow -\frac{\partial}{\partial \theta} [\log p(\theta|\psi_t, \mathbf{y})]|_{\theta=\theta_l^*};$
 $\mathbf{p}_l^* \leftarrow \mathbf{p}_{l,1}^* - \frac{\epsilon}{2} \mathbf{g}_2;$
end

/* Calculate Hamiltonian (total energy) of initial position */
 $K_{t-1} \leftarrow \frac{1}{2}(\mathbf{p}_0^*)^T M^{-1} \mathbf{p}_0^*;$
 $U_{t-1} \leftarrow -\log p(\theta_0^*|\psi_t, \mathbf{y});$
 $H_{t-1} \leftarrow K_{t-1} + U_{t-1};$

/* Calculate Hamiltonian (total energy) of candidate position */
 $K^* \leftarrow \frac{1}{2}(\mathbf{p}_L^*)^T M^{-1} \mathbf{p}_L^*;$
 $U^* \leftarrow -\log p(\theta_L^*|\psi_t, \mathbf{y});$
 $H^* \leftarrow K^* + U^*;$

/* Metropolis correction to determine if proposal accepted */
Draw $u \sim \text{Unif}[0, 1];$
 $\log r \leftarrow H_{t-1} - H^*;$
if $\log u < \log r$ **then**
 $\theta_t \leftarrow \theta_L^*$
else
 $\theta_t \leftarrow \theta_{t-1}$
end

Since the likelihood is a function of β_f , we need to use the chain rule to get the gradient in μ_f space:

$$\begin{aligned} \frac{\partial}{\partial \mu_f} \left[\log p(\mu_f | \mathbf{W}, \mathbf{I}, \mathbf{l}, \{\tau_f^2\}_{f=1}^V, \psi, \gamma^2, \{\xi_d\}_{d=1}^D, \mathcal{T}) \right] \\ = \frac{\partial l(\beta_f)}{\partial \beta_f} \frac{\partial \beta_f}{\partial \mu_f} + \frac{\partial}{\partial \mu_f} \left[\log p(\mu_f | \{\tau_f^2\}_{f=1}^V, \psi, \gamma^2, \mathcal{T}) \right] \\ = - \sum_{d=1}^D l_d(\theta_d^T \circ \beta_f^T) + \sum_{d=1}^D \left(\frac{w_{fd}}{\theta_d^T \beta_f} \right) (\theta_d^T \circ \beta_f^T) - \Lambda(\mu_f - \psi \mathbf{1}), \end{aligned}$$

where \circ is the Hadamard (entrywise) product. The Hessian matrix follows a similar pattern:

$$\mathbf{H}(\log p(\mu_f | \mathbf{W}, \mathbf{I}, \mathbf{l}, \{\tau_f^2\}_{f=1}^V, \psi, \gamma^2, \{\xi_d\}_{d=1}^D, \mathcal{T})) = -\Theta^T \mathbf{W} \Theta \circ \beta_f \beta_f^T + \mathbf{G} - \Lambda,$$

where

$$\mathbf{W} = \text{diag} \left(\left\{ \frac{w_{fd}}{(\theta_d^T \beta_f)^2} \right\}_{d=1}^D \right)$$

and

$$\mathbf{G} = \text{diag} \left(\frac{\partial l(\beta_f)}{\partial \beta_f} \circ \beta_f^T \right) = \text{diag} \left(\frac{\partial l(\beta_f)}{\partial \mu_f} \right).$$

We use the BFGS algorithm with the analytical gradient derived above to maximize this density for iterations where the likelihood Hessian is updated; this quasi-Newton method works well since the conditional posterior is unimodal. The Hessian of the likelihood in β space is clearly negative definite everywhere since $\Theta^T \mathbf{W} \Theta$ is a positive definite matrix. The prior Hessian Λ is also positive definite by definition since it is the precision matrix of a Gaussian variate. However, the contribution of the chain rule term \mathbf{G} can cause the Hessian to become indefinite away from the mode in μ space if any of the gradient entries are sufficiently large and positive. Note, however, that the conditional posterior is still unimodal since the logarithm is a monotone transformation.

A.2.2 Conditional posterior of the topic affinity parameters

The log conditional posterior for the topic affinity parameters for one document is:

$$\begin{aligned}
\log p(\boldsymbol{\xi}_d | \mathbf{W}, \mathbf{I}, \mathbf{l}, \{\boldsymbol{\mu}_f, \boldsymbol{\tau}_f^2\}_{f=1}^V, \boldsymbol{\eta}, \boldsymbol{\Sigma}) \\
&= l_d \sum_{f=1}^V \log \text{Pois}(w_{fd} | \boldsymbol{\beta}_f^T \boldsymbol{\theta}_d) + \log \text{Bernoulli}(\mathbf{I}_d | \boldsymbol{\xi}_d) + \log \mathcal{N}(\boldsymbol{\xi}_d | \boldsymbol{\eta}, \boldsymbol{\Sigma}) \\
&= -l_d \sum_{f=1}^V \boldsymbol{\beta}_f^T \boldsymbol{\theta}_d + \sum_{f=1}^V w_{fd} \log(\boldsymbol{\beta}_f^T \boldsymbol{\theta}_d) - \sum_{k=1}^K \log(1 + \exp(-\xi_{dk})) \\
&\quad - \sum_{k=1}^K (1 - I_{dk}) \xi_{dk} - \frac{1}{2} (\boldsymbol{\xi}_d - \boldsymbol{\eta})^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{\xi}_d - \boldsymbol{\eta}).
\end{aligned}$$

Since the likelihood of the word counts is a function of $\boldsymbol{\theta}_d$, we need to use the chain rule to get the gradient of the likelihood in $\boldsymbol{\xi}_d$ space. This mapping is more complicated than in the case of the $\boldsymbol{\mu}_f$ parameters since each ξ_{dk} is a function of all elements of $\boldsymbol{\theta}_d$:

$$\nabla l_d(\boldsymbol{\xi}_d) = \nabla l_d(\boldsymbol{\theta}_d)^T \mathbf{J}(\boldsymbol{\theta}_d \rightarrow \boldsymbol{\xi}_d),$$

where $\mathbf{J}(\boldsymbol{\theta}_d \rightarrow \boldsymbol{\xi}_d)$ is the Jacobian of the transformation from $\boldsymbol{\theta}$ space to $\boldsymbol{\xi}$ space, a $K \times K$ symmetric matrix. Let $S = \sum_{l=1}^K \exp \xi_{dl}$. Then

$$\mathbf{J}(\boldsymbol{\theta}_d \rightarrow \boldsymbol{\xi}_d) = S^{-2} \begin{bmatrix} S \exp \xi_{d1} - \exp 2\xi_{d1} & \dots & -\exp(\xi_{dK} + \xi_{d1}) \\ -\exp(\xi_{d1} + \xi_{d2}) & \dots & -\exp(\xi_{dK} + \xi_{d2}) \\ \vdots & \ddots & \vdots \\ -\exp(\xi_{d1} + \xi_{dK}) & \dots & S \exp \xi_{dK} - \exp 2\xi_{dK} \end{bmatrix}.$$

The gradient of the likelihood of the word counts in terms of $\boldsymbol{\theta}_d$ is

$$\nabla l_d(\boldsymbol{\theta}_d) = -l_d \sum_{f=1}^V \boldsymbol{\beta}_f^T + \sum_{f=1}^V \frac{w_{fd} \boldsymbol{\beta}_f^T}{\boldsymbol{\beta}_f^T \boldsymbol{\theta}_d}.$$

Finally, to get the gradient of the full conditional posterior, we add the gradient of the likelihood of the labels and of the normal prior on the $\boldsymbol{\xi}_d$:

$$\begin{aligned}
\frac{\partial}{\partial \boldsymbol{\xi}_d} \left[\log p(\boldsymbol{\xi}_d | \mathbf{W}, \mathbf{I}, \mathbf{l}, \{\boldsymbol{\mu}_f\}_{f=1}^V, \boldsymbol{\eta}, \boldsymbol{\Sigma}) \right] \\
= \nabla l_d(\boldsymbol{\theta}_d)^T \mathbf{J}(\boldsymbol{\theta}_d \rightarrow \boldsymbol{\xi}_d) + (\mathbf{1} + \exp \boldsymbol{\xi}_d)^{-1} - (\mathbf{1} - \mathbf{I}_d) - \boldsymbol{\Sigma}^{-1} (\boldsymbol{\xi}_d - \boldsymbol{\eta}).
\end{aligned}$$

The Hessian matrix of the conditional posterior is a complicated tensor product that is not efficient to evaluate analytically. Instead, we compute a numerical Hessian using the analytic gradient presented above at minimal computational cost.

We use the BFGS algorithm with the analytical gradient derived above to maximize this density for iterations where the likelihood Hessian is updated. We have not been able to show analytically that this conditional posterior is unimodal, but we have verified this graphically for several documents and have achieved very high acceptance rates for our HMC proposals based on this Hessian calculation.

A.2.3 Conditional posterior of the τ_{fk}^2 hyperparameters

The variance parameters τ_{fk}^2 independently follow an identical Scaled Inverse- χ^2 with convolution parameter ν and scale parameter σ^2 , while their inverse follows a Gamma($\kappa_\tau = \frac{\nu}{2}, \lambda_\tau = \frac{2}{\nu\sigma^2}$) distribution. The log conditional posterior of these parameters is:

$$\begin{aligned} \log p(\kappa_\tau, \lambda_\tau | \{\tau_f^2\}_{f=1}^V, \mathcal{T}) &= (\kappa_\tau - 1) \sum_{f=1}^V \sum_{k \in \mathcal{P}} \log (\tau_{fk}^2)^{-1} \\ &\quad - |\mathcal{P}|V\kappa_\tau \log \lambda_\tau - |\mathcal{P}|V \log \Gamma(\kappa_\tau) - \frac{1}{\lambda_\tau} \sum_{f=1}^V \sum_{k \in \mathcal{P}} (\tau_{fk}^2)^{-1}, \end{aligned}$$

where $\mathcal{P}(\mathcal{T})$ is the set of parent topics on the tree. If we allow $i \in \{1, \dots, N = |\mathcal{P}|V\}$ to index all the f, k pairs and $l(\kappa_\tau, \lambda_\tau) = p(\{\tau_f^2\}_{f=1}^V | \kappa_\tau, \lambda_\tau, \mathcal{T})$, we can simplify this to

$$l(\kappa_\tau, \lambda_\tau) = (\kappa_\tau - 1) \sum_{i=1}^N \log \tau_i^{-2} - N\kappa_\tau \log \lambda_\tau - N \log \Gamma(\kappa_\tau) - \frac{1}{\lambda_\tau} \sum_{i=1}^N \tau_i^{-2}.$$

We then transform this density onto the $(\log \kappa_\tau, \log \lambda_\tau)$ scale so that the parameters are unconstrained, a requirement for standard HMC implementation. Each draw of $(\log \kappa_\tau, \log \lambda_\tau)$ is then transformed back to the (ν, σ^2) scale. To get the Hessian of the likelihood in log space, we calculate the derivatives of the likelihood in the original space and apply the chain rule:

$$\begin{aligned} \mathbf{H} \left(l(\log \kappa_\tau, \log \lambda_\tau) \right) &= \\ &= \begin{bmatrix} \kappa_\tau \frac{\partial l(\kappa_\tau, \lambda_\tau)}{\partial \kappa_\tau} + (\kappa_\tau)^2 \frac{\partial^2 l(\kappa_\tau, \lambda_\tau)}{\partial (\kappa_\tau)^2} & \kappa_\tau \lambda_\tau \frac{\partial^2 l(\kappa_\tau, \lambda_\tau)}{\partial \kappa_\tau \partial \lambda_\tau} \\ \kappa_\tau \lambda_\tau \frac{\partial^2 l(\kappa_\tau, \lambda_\tau)}{\partial \kappa_\tau \partial \lambda_\tau} & \lambda_\tau \frac{\partial l(\kappa_\tau, \lambda_\tau)}{\partial \lambda_\tau} + (\lambda_\tau)^2 \frac{\partial^2 l(\kappa_\tau, \lambda_\tau)}{\partial (\lambda_\tau)^2} \end{bmatrix}, \end{aligned}$$

where

$$\nabla l(\kappa_\tau, \lambda_\tau) = \begin{bmatrix} \sum_{i=1}^N \log \tau_i^{-2} - N \log \lambda_\tau - N\psi(\kappa_\tau) \\ -\frac{N\kappa_\tau}{\lambda_\tau} + \frac{1}{(\lambda_\tau)^2} \sum_{i=1}^N \tau_i^{-2} \end{bmatrix}$$

and

$$\mathbf{H} \left(l(\kappa_\tau, \lambda_\tau) \right) = \begin{bmatrix} -N\psi'(\kappa_\tau) & -\frac{N}{\lambda_\tau} \\ -\frac{N}{\lambda_\tau} & \frac{N\kappa_\tau}{(\lambda_\tau)^2} - \frac{2}{(\lambda_\tau)^3} \sum_{i=1}^N \tau_i^{-2} \end{bmatrix}.$$

Following Algorithm 1, we evaluate the Hessian at the mode of this joint posterior. This is easiest to find on original scale following the properties of the Gamma distribution. The first order condition for λ_τ can be solved analytically:

$$\lambda_{\tau,MLE}(\kappa_\tau) = \arg \max_{\lambda_\tau} \left\{ l(\kappa_\tau, \lambda_\tau) \right\} = \frac{1}{\kappa_\tau N} \sum_{i=1}^N \tau_i^{-2}.$$

We can then numerically maximize the profile likelihood of κ_τ :

$$\kappa_{\tau,MLE} = \arg \max_{\kappa_\tau} \left\{ l(\kappa_\tau, \lambda_{\tau,MLE}(\kappa_\tau)) \right\}.$$

The joint mode in the original space is then $(\kappa_{\tau,MLE}, \lambda_{\tau,MLE}(\kappa_{\tau,MLE}))$. Due to the monotonicity of the logarithm function, the mode in the transformed space is simply $(\log \kappa_{\tau,MLE}, \log \lambda_{\tau,MLE})$. We can be confident that the conditional posterior is unimodal: the Fisher information for a Gamma distribution is negative definite, and the log transformation to the unconstrained space is monotonic.